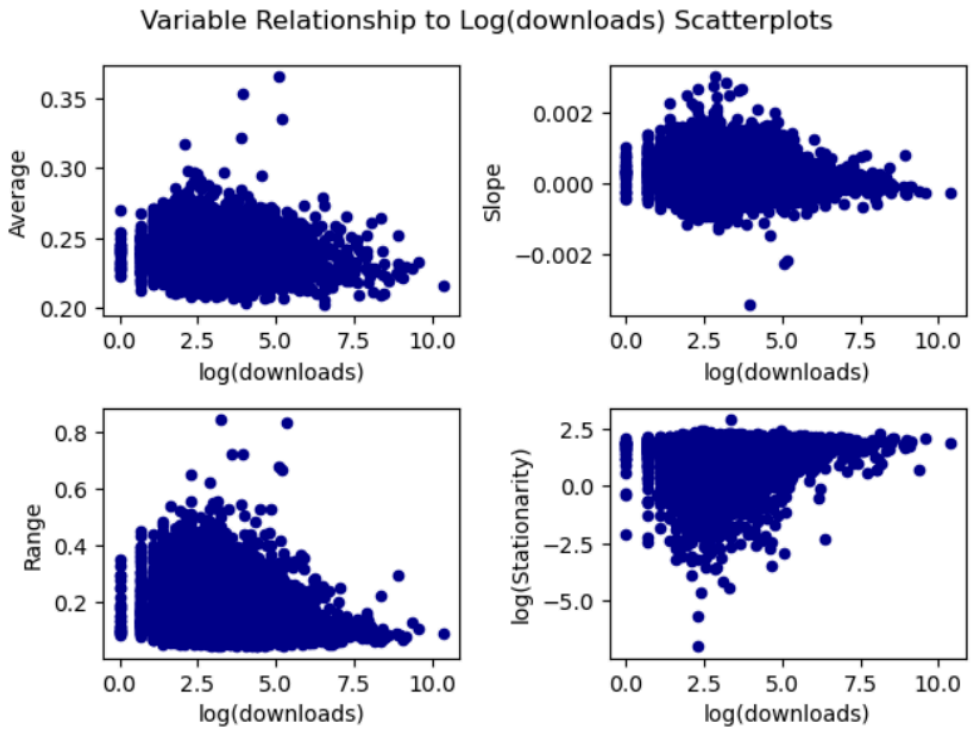
Andrew Nee

andrewnee29@gmail.com

Information Revelation

Information revelation undoubtedly has a role in the popularity of books. The book-level measures of characteristics of the Kullback-Liebler divergence I implemented were Average, Variance, Slope, Min, Max, Range, and Stationarity. I then plotted some variable relationships to log(downloads).



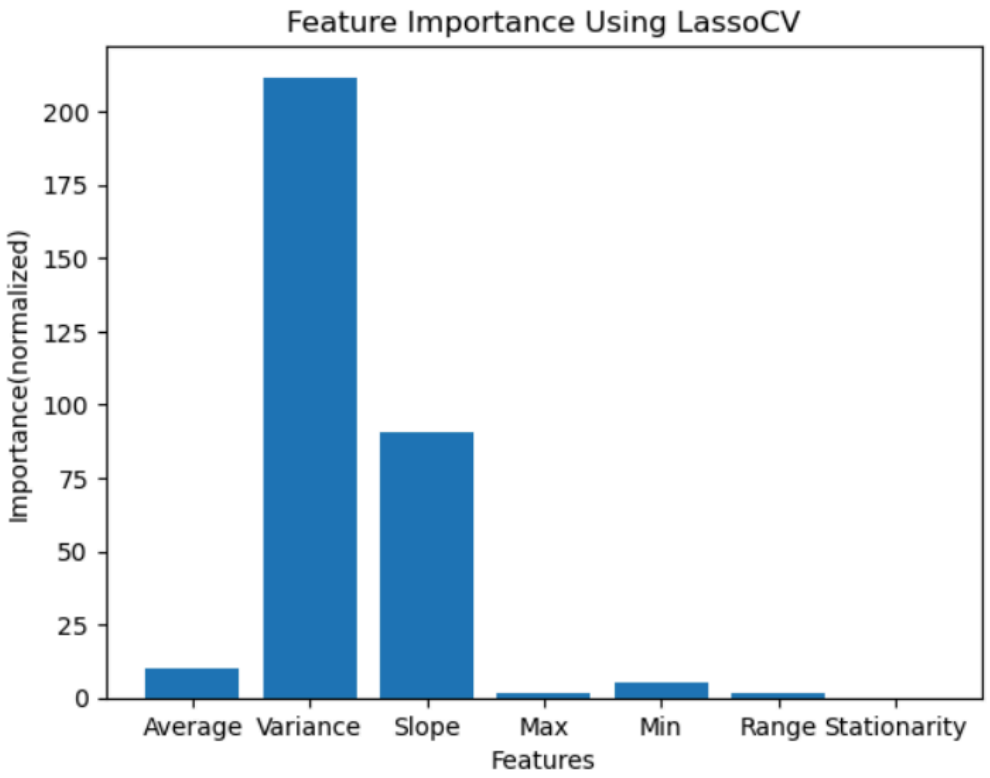
There seems to be an arrowhead-like relationship for all of the variables. As log(downloads) increases, Average converges towards ~0.23, Slope to ~0.000, Range to ~.1, and log(Stationarity) to ~2.4. I was particularly interested in the attributes that represented consistency such as Slope, Range, and Stationarity, which treats the series of information relevance as time-series data and measures stationarity using an augmented Dickey-Fuller test. In this case, I took the absolute value of the test in order to take the log of the data, so the higher the

number, the more stationary the data is. The data points towards popular books being more consistent, meaning that the range of information relevance is lower, the slope is flatter, and stationarity is higher.

	index	Romance	Fantasy	SciFi	Adventure	History
0	Average	0.236617	0.241597	0.243643	0.237161	0.236294
1	Variance	0.001162	0.001065	0.001063	0.001055	0.000978
2	Slope	0.000334	0.000235	0.000290	0.000302	0.000277
3	Max	0.374947	0.360201	0.370237	0.365488	0.356615
4	Min	0.194022	0.196953	0.199700	0.194323	0.195307
5	Range	0.180925	0.163248	0.170537	0.171165	0.161309
6	Stationarity	4.715743	5.071877	5.007202	4.848869	4.867488
7	downloads	217.703529	211.421053	125.252336	99.130068	77.660042
8	log(downloads)	3.247432	4.246834	4.197575	3.067011	3.072446
9	log(Stationarity)	1.352022	1.475285	1.452990	1.405530	1.411648

Across multiple English-fiction genres, the features didn't drastically vary from each other, and were rather homogeneous. It would be interesting to contrast this data to English-nonfiction genres, such as biographies, to investigate information relevance characteristics.

Lastly, I train-test-split and normalized the data to fit a LassoCV model to get feature importance. The model found Variance and Slope to be the most independently predictive of log(downloads).



Tables:

	Average	Variance	Slope	Max	Min	Range	Stationarity	downloads	log(downloads)
0	0.234033	0.001644	0.000870	0.450747	0.192720	0.258026	0.099532	593.0	6.385194
1	0.243351	0.001105	0.000143	0.390158	0.206327	0.183832	5.622344	17.0	2.833213
2	0.240153	0.002246	0.000813	0.548940	0.204162	0.344778	1.015496	47.0	3.850148
3	0.247845	0.001715	0.000090	0.481949	0.203087	0.278862	7.468118	19.0	2.944439
4	0.250666	0.001194	-0.000112	0.391893	0.201637	0.190256	6.093013	12.0	2.484907
...	...	...	...	...	...	...	...	...	...
7758	0.238457	0.000984	0.000811	0.336083	0.185911	0.150171	4.170598	14.0	2.639057
7759	0.235281	0.001369	0.000817	0.459716	0.203732	0.255983	0.385642	84.0	4.430817
7760	0.234510	0.000634	0.000134	0.338172	0.192124	0.146047	6.184309	8.0	2.079442
7761	0.228159	0.000825	0.000612	0.406871	0.196882	0.209988	5.211601	28.0	3.332205
7762	0.208989	0.000190	-0.000162	0.252348	0.189248	0.063100	6.771437	4257.0	8.356320

Attributes:

Average - using np.mean to get the mean of the kld_values series

Variance - np.var

Slope - created a calculate_slope() function that utilized linregress() to find the slope of the series

Max - np.max

Min - np.min

Range - (Max - Min)

Stationarity - created a calculate_adf() function that utilized adfuller() to find the absolute value of the test statistic